



CLARIN ERIC and CLARIN-IT

Francesca Frontini - *CLARIN Board of Directors*

Monica Monachini - CLARIN-IT National Coordinator

Istituto di Linguistica Computazionale - ILC CNR

13/06/2023

Links in Presentation

Open Data, SSH, Research Infrastructures

- ELTeC project: <https://www.distant-reading.net/>

Schöch, Christof, Erjavec, Tomaz, Patras, Roxana, & Santos, Diana. (2021). Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives. Modern Languages Open. <https://doi.org/10.5281/zenodo.4742420>

Carolin Odebrecht, Lou Burnard, & Christof Schöch. (2021). European Literary Text Collection (ELTeC): April 2021 release with 14 collections of at least 50 novels. (v1.1.0) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.4662444>

- DRA COR: <https://dracor.org/>

Fischer, Frank, et al. (2019). Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. In Proceedings of DH2019: "Complexities", Utrecht University, <https://doi.org/10.5281/zenodo.4284002>.

- Programming Historian: <https://programminghistorian.org/>

- European Commission, Directorate-General for Research and Innovation, Open innovation, open science, open to the world : a vision for Europe, Publications Office, 2016, <https://data.europa.eu/doi/10.2777/061652>

- OECD, Open Science. <https://www.oecd.org/sti/inno/open-science.htm>

- FOSTER, Open Science at the Core of Libraries. <https://www.fosteropenscience.eu/learning/open-science-at-the-core-of-libraries>

- The Open Science and Research Handbook v1 https://cdn2.euraxess.org/sites/default/files/open_science_and_research_handbook_v.1.0.pdf

- 'FAIR Guiding Principles for scientific data management and stewardship', 2016 <https://www.go-fair.org/fair-principles/>

- <https://www.openaire.eu/what-is-a-data-management-plan>

How to use the CLARIN infrastructure for SSH research

- SHAPE-ID. (2021). SHAPE-ID Toolkit Resources - Research Infrastructures Collection. Zenodo. <https://doi.org/10.5281/zenodo.5116029>
<https://www.shapeidtoolkit.eu/wp-content/uploads/2021/03/Guide-AHSS-Research-Infrastructures.pdf>
- <https://www.clarin.eu/>
- <https://www.dariah.eu/>
- <https://www.clarin.eu/content/overview-clarin-centres>
- LINDAT CLARIAH
 - Portal <https://lindat.cz/>
 - Repository <https://lindat.mff.cuni.cz/repository/xmlui/?locale-attribute=en>
 - Tools <https://lindat.cz/#tools>
- CLARIN-UK <https://www.clarin.ac.uk/>
- CLARIN ERIC Portal: <https://www.clarin.eu/>
 - CLARIN supporting FAIR <https://www.clarin.eu/fair>
 - Virtual Language Observatory (VLO) <https://vlo.clarin.eu/>
 - Language Resources Switchboard (LRS) <https://switchboard.clarin.eu/>
- Riesner, Katherina (2017). The Barack Obama Corpus [Data set]. <http://hdl.handle.net/10050/00-0000-0000-0003-C53B-4@view>
- Straka, Milan and Straková, Jana, 2016, UDPipe, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, <http://hdl.handle.net/11234/1-1702>
- CLARIN RESOURCE FAMILIES: <https://www.clarin.eu/resource-families>
- CLARIN and OPEN and FAIR Science
 - CLARIN: [Towards FAIR and Responsible Data Science Using Language Resources.](#) In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), May 2018, 3259-3264.

CLARIN for Knowledge Sharing

- <https://www.clarin.eu/content/knowledge-infrastructure>
- <https://www.clarin.eu/content/knowledge-centres>
- DH Course Registry
<https://dhcr.clarin-dariah.eu/>

- <https://www.clarin.eu/Tour-de-CLARIN>
- <https://www.clarin.eu/content/clarin-impact-stories>

- <https://www.clarin.eu/content/teaching-clarin>
- <https://www.clarin.eu/content/clarin-cafe>

- FUNDING: <https://www.clarin.eu/content/funding-opportunities>

- The results of the Helsinki Hackathon (2021)
<https://www.clarin.eu/impact-stories/helsinki-digital-humanities-hackathon-2021-parliamentary-debates-covid-times>

The CLARIN-IT National Consortium

- CLARIN-IT national Consortium
<https://www.clarin-it.it/it>
- ILC4CLAIN B centre and Repository
 - <https://ilc4clarin.ilc.cnr.it/>
 - <https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/>

A Cluster of SSH Research Infrastructures

- <https://eosc-portal.eu/>
- <https://marketplace.eosc-portal.eu/>
- <https://www.sshopencloud.eu/>
- <https://marketplace.sshopencloud.eu/>

Exercises

Exercise 1 - Data Citation

Tromsø recommendations for citation of research data in linguistics.

[DOI: 10.15497/rda00040](https://doi.org/10.15497/rda00040)

See also: <https://youtu.be/GyBCslbn6tc>

Recommended fields (the fields in *Italics* are optional):

- Author (main investigator)
- Other attribution roles (e.g. Data collector)
- Date (date of publication or deposit)
- Title (name of the resource)
- Publisher (data repository or organisation)
- Locator (DOI, URN or URL)
- Version: References to versioned datasets should include the version number.

Check the following citations and assess their adherence to these recommendations

- Fabián Villena. (2019). Multilingual Medical Corpora [Data set]. In *Studies in Health Technology and Informatics: Vol. Volume 270: Digital Personalized Health and Medicine* (pp. 347–351). Zenodo. <https://doi.org/10.5281/zenodo.3463379>
- Hajič, Jan; et al., 2020, Prague Dependency Treebank - Consolidated 1.0 (PDT-C 1.0), LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, <http://hdl.handle.net/11234/1-3185>.
- Brezina, V., Weill-Tessier, P., & McEnery, A. (2020). #LancsBox v. 5.x. [software]. Available at: <http://corpora.lancs.ac.uk/lancsbox>.
- Davies, Mark. *The Corpus of Contemporary American English*. 2008, www.english-corpora.org/coca/.
- BNC Consortium. *The British National Corpus, version 3 (BNC XML Edition)*. 2007. Oxford: Bodleian Libraries, University of Oxford. [http:// www.natcorp.ox.ac.uk/](http://www.natcorp.ox.ac.uk/).
- Huang, Shudong, David Graff and George Doddington. *Multiple-Translation Chinese Corpus LDC2002T01*. Web download file. Philadelphia: Linguistic Data Consortium, 2002.
- Princeton University “About WordNet.” WordNet. Princeton University. 2010.

<https://www.rd-alliance.org/groups/data-citation-wg.html>

Exercise 2 - Finding and processing data

- Go to clarin.eu

- Find the VLO on the CLARIN website (tip: it is a service that allows you to search for language resources...)
 - Take a look at the examples of queries
- Search for texts by Robert Louis Stevenson
 - What can you find?
 - Which format is the corpus in?
 - Where are they hosted?
- On the Links tab, use the three dots (...) to activate the Switchboard (see image)
 - Explore with **Voyant** or
 - Process with **UDpipe** <https://lindat.mff.cuni.cz/services/udpipe/>
- You can watch the CLARIN video to get an overall idea of CLARIN's services
 - <https://www.clarin.eu/content/about-clarin>

The screenshot shows the VLO interface with the following elements:

- Header: Virtual Language Observatory, Search, Contributors, Help, CLARIN logo.
- Breadcrumbs: VLO / Faceted search / Search results / Record: Treasure Island / Robert Louis Stevenson.
- Record 1 of 36.
- Navigation: < previous, next >
- Record title: Treasure Island / Robert Louis Stevenson.
- Record details tabs: Record details (selected), Links (5), Availability, All metadata, Technical Details.
- Table of links:

Name	Type
HDL 2055	landing page
dublin_core.xml	XML
metadata_local.xml	XML
header2055.xml	XML
treas10-2055.bt	Plain Text

At the bottom of the table, there is a button: Process with Language Resource Switchboard.

Footer: About v4.10.2, Service provided by CLARIN, Contact.

Exercise 3 - The ParlaMint corpora

More detailed instructions here:

Darja Fišer, Kristina Pahor 2021. de Maiti Voices of the Parliament : A Corpus Approach to Parliamentary Discourse Research <http://hdl.handle.net/20.500.12325/121>

See also <https://sidih.github.io/voices/index.html>

- Explore the ParlaMint corpora using NoSketch Engine (public): <https://clarin.si/noske/index-en.html>
- Choose a corpus (Italian Senate, UK Parliament)
- Explore the **Corpus info** -

- How many subcorpora are there?
- What are the attributes annotated for Speech?

Home

Search

Word list

Corpus info

My jobs

User guide [↗](#)

Menu position

ParlaMint-IT 2.1 (Italian parliament) [?](#)

Italian parliamentary corpus ParlaMint-IT, 2013-2020 v2.1

Counts		General info		Lexicon sizes	
Tokens	30,615,130	Corpus description	Document	word	171,465
Words	26,571,966	Language	Italian	lc ?	156,333
Sentences	1,087,465	Encoding	UTF-8	norm ?	175,208
Paragraphs	214,300	Compiled	06/16/2021 18:42:31	lemma	110,155
Documents	79,283	Tagset	Description	lemma_lc ?	105,563
				pos ?	67
				feats ?	803
				n ?	2,148
				dep ?	310
				dep_head_lemma ?	142,174
				dep_head_pos ?	146
				dep_head_feats ?	1,707
				dep_head_n ?	6,311

Structures and attributes

name	868,371	?
note	116,053	?
p	214,300	?
s	1,087,465	?
speech	79,283	?

Subcorpora statistics

Subcorpus	Tokens	Words	%
BiL_Stab_Manov	8,500,237	- 7,377,659	27.76
Bilancio	6,252,190	- 5,426,499	20.42
COVIDMPMALE	2,095,859	- 1,819,070	6.84
Coalition	7,010,950	- 6,085,061	22.90

- Create a wordlist for the COVID subcorpus
- Perform Keyword extraction
 - extract keywords for the “COVID” subcorpus using the rest of the corpus as Reference (sub)corpus

Corpus: ParlaMint-IT 2.1 (Italian parliament)

Subcorpus: Covid [info](#) [create new](#) ?

Search attribute: lemma

use n-grams. Value of n: from 2 to 2 ?

hide/nest sub-n-grams

Filter options:

Filter word list by: Regular expression:

Minimum frequency: 5

Maximum frequency: 0 (0 = no maximum frequency)

Whitelist: No file chosen

Blacklist: No file chosen [format](#)

Include non-words

Output options:

Frequency figures: Hit counts Document counts ARF

Output type: Simple Keywords

Reference (sub)corpus: ParlaMint-IT 2.1 (Italian parliament) the rest of the corpus


Prefer: rare words common words 1

Change output attribute(s)

You can select one or more output attributes. Please note that this option can be time-consuming.

Exercise 4 - Explore ILC4CLARIN

- Find the [ILC4CLARIN repository](#)
 - Which is the most represented language in terms of records?
 - What kind of data can you find in Arabic?
 - How do you cite CophiWordNet?
- Try to log in with your institutional identifier, using the Login function (top right)

 Sign in via the CLARIN Service Provider Federation



Select your home organisation below. This is usually the organisation where you work or study. Signing in here will allow you to access certain CLARIN resources and services which are only available to users who have logged in. If you cannot find your organisation in the list below, please select the clarin.eu website account and use your CLARIN website credentials. If you don't have such credentials you can register an account [here](#). For questions please contact spf@clarin.eu.

Previously chosen home organisation

CNR Institute for Computational Linguistics "Antonio Zampolli"
Italy

ILC

Home organisation list

All countries



clarin.eu website account
European Union



AAI@EduHr Single Sign-On Service
Croatia



Aalborg University
Denmark

Exercise 5 - The SSH Open Marketplace

- Explore the training materials on the [SSHOC Marketplance](#)
- What is the difference between the SSH Open Marketplace and the CLARIN VLO?
- Can you find any of the datasets I have cited at the beginning of the presentation?
- What are workflows? What is the relationship between workflows and other classes of objects?

Further reading and references

Training Materials

Yankelevich, Tanya, Fiser, Darja, Lenardic, Jakob, Gorgaini, Elisa, & Braukmann, Ricarda. (2020, June). LIBER 2020 - Workshop: SSHOC Train-the-Trainer Bootcamp for Librarians. Zenodo. <http://doi.org/10.5281/zenodo.3970799>

Darja Fišer, Kristina Pahor de Maiti 2021. Voices of the Parliament : A Corpus Approach to Parliamentary Discourse Research <http://hdl.handle.net/20.500.12325/121>

Del Fante, Dario. (2022). ParlaMint – IT – Il corpus del Senato Italiano. Una guida pratica per l'interrogazione del corpus ParlaMint-IT con NoSketch Engine, a supporto dell'analisi del discorso politico. Zenodo. <https://doi.org/10.5281/zenodo.6526914>

Francesca Frontini, Andrea Bellandi, Valeria Quochi, Monica Monachini, Karlheinz Mörth, Susanne Zhanial, Matej Ďurčo and Anna Woldrich (2022). CLARIN Tools and Resources for Lexicographic Work. Version 1.0.0. DARIAH-Campus. [Training module]. <https://elexis.humanistika.org/id/UnwYPq70Dewbn7XDEjsMM>

Keeping in Touch

- Visit this page
<https://www.clarin.eu/content/clarin-researchers>
- Participate in CLARIN (virtual) events
<https://www.clarin.eu/events>
- Subscribe to the NewsFlash
<https://www.clarin.eu/news>
- Follow CLARIN and CLARIN-IT on Twitter
@CLARINERIC

- EMAIL US: coordination@clarin-it.it